Electronic Version
Stylesheet Version v1.1.1

Claims

[c1] A method comprising the steps of:

a) Performing Level 1: Indexing / Classification applied to data corpus "A", where "A" is a data corpus consisting of (typically) a large to very large number of members which can be structured, semi-structured, and/or un-structured text, the result(s) of any form of speech-to-text conversion, and/or images or other signal-processed data, and/or any combination of such data, where the Indexing / Classification process is performed specifically as: indexing and /or classifying the members of data corpus "A"by appending to each member one or more "metatags" descriptive of the content of that mem-ber, whether that content is explicitly referenced (e.g., via "indexing," using methods and terminology well known to practitioners of the art), or implicitly refer-enced using one or more of the various possible "classifi-cation" algorithms (e.g., Bayesian, or Bayesian aug-mented with "Shannon Information Theory" feature vec-tor weighting), where the only specific requirement of the classification algorithm(s) at least one of the algo-rithm(s) employed be "controllable" through at least one parameter value (e.g., the "sigma" value in a Bayesian

classifier, or more broadly, the "sigma" value, the number of elements in the prototyping "feature vector" for such a classifier, and the "feature vector element weights" applied to each element of a given "feature vector," where these terms and associated methods are all well known to practitioners of the art, and this specification of possible parameter types is by no means exhaustive), and the end result is the set of one or more "metatags" so produced by application of one or more classification algorithm(s) to a given data corpus element and then associated with that element are indicative of the content of each element; and additionally a document may be classified and / or metatagged as containing one or more concept classes whose existence is inferred through the presence of certain words (typically noted as feature vectors) in that document,

b) Performing Level 1 to Level 2 Transition, by which a proper subset of members from the initial data corpus "A" are selected for Level 2 processing, which is done by selecting from among all the (optionally indexed and) metatagged members of data corpus "A" those whose metatags are a match to a set of criteria, where these criteria can be set either or both by the user of this method or by an automated process incorporated as part of this method, and whose exact specification does not in any way impact the generality of the method de-

scribed here, and this subset is denoted data corpus "B",
c) Performing Level 2 Pairwise Associative Processing, by which the data corpus "B" members selected during said step (b) are processed so as to produce "pairwise associations" between the elements of each of these members of "B", where a typical embodiment of this step would be to generate a set of pairwise associations of nouns and / or noun phrases extracted from a text-based corpus "B", although this method can be extended and applied to data corpora containing other types of elements (e.g. images, signals) without loss of meaning or generality, and where the associations are typically limited to those within a given member of "B", although the results of such associations are typically noted accumulatively across the entire corpus "B", and a typical embodiment of this step is a "pairwise co-occurrence matrix" applied to objects in each member of "B" whereby a corresponding matrix element is incremented whenever a given pair of nouns and / or noun phrases occurs within a set distance of each other, although any accumulative pairwise-association method applied across "B" may be used without loss of the generality or meaning of the knowledge discovery method being described herein,

[c2] A method as claimed in c1 further comprising the optional step of:

a) Performing Level 0: Optional Preprocessing / Indexing, specifically: (optionally) indexing the members of a data corpus "A0" by "tagging" each member of the corpus with one or more "metatags" in any such manner as is well known to practitioners of the art, whereby the "metatags" refer to specific identifiable elements (e.g., but not limited to, specific words, or specific content as might be found in an image) and where this step is typically reserved for very large data corpora (e.g., typically where the number of members of data corpus "A0" exceeds $O(10^6)$) but may be applied to any size corpus without loss of the validity or generality of this method;
b) Performing Level 0 to Level 1 Transition, specifically selecting those members of the data corpus whose "indices" as found and applied in said step (a) are a "match" to some specified criteria, whether these criteria are set manually by a user for a given knowledge discovery task or set via an automated process, and the method by which these "index matches" are selected is any one of those well known to practitioners of the art and detailed specification of such method or development of a new "indexing" method is not essential to specifying this knowledge discovery method, nor is it essential to specify the method by which such "indexed" data corpus members are "selected" for "Transition" to the predecessor step (1a) except that the general intention of said "se-

lection" is to reduce the size of the "selected" sub-corpus, which we now denote corpus "A".

[c3] A method as claimed in either c1 or c2, further comprising the step of:

a) Performing Level 2 to Level 3 Transition, by which the "pairwise associations" found in said step (1c) are filtered by any one or more of various algorithmic means well known to the practitioners of this art so as to extract a subset of associations by application of one or more selection criteria, and the generality and meaning of this method is not dependent upon the specific nature of these criteria, and where a typical embodiment of this method would be to use a cut-off process selecting only those "pairwise associations" that reach a certain predefined or preset value, whether this value is fixed or determined by an algorithmic means (such as histogramming or thresholding, or any such method as is employed by the community for similar purposes), and where extracted subset of these associations is hereafter referred to as data corpus "C" and is passed to a subsequent "Level 3" for further processing,

b) Performing Level 3 Syntactic Associative Processing, by which the data corpus "C" members selected during said step (3a) are processed so as to produce "syntactic associations" between the elements of one or more of

each of these members of "C", where a typical embodiment of this step would be to generate a set of subject noun-verb-object noun associations using nouns and / or noun phrases extracted from the data corpus "C" as subject nouns (and potentially also as object nouns) and the verbs and additional object nouns are drawn from the data sources from which data corpus "B" was extracted, although this method can also include simple subject noun-verb associations and also verb-object noun associations, and where the identifications of subject nouns, object nouns, noun phrases, concept classes, and verbs, are those common to practitioners of the art, and the resulting representation of the syntactically-associated may be either in structured (e.g., database) or other form, so long as the syntactic relationship between the associated words or phrases is represented, and may also include, without loss of generality or meaning of this method, additional grammatical annotations to the basic syntactic representation (e.g., adjectives, etc.) and any one or more noun and / or noun phrase may be replaced with an associated "concept class, "using methods that are the same or similar to those described in (1a),

[c4] A method as claimed in c3, further comprising the step of:

a) Performing Level 3 to Level 4 Transition, by which the "syntactic associations" found in said step (3b) are filtered by any one or more of various algorithmic means well known to the practitioners of this art so as to extract a subset of associations by application of one or more selection criteria, and the generality and meaning of this method is not dependent upon the specific nature of these criteria, and this subset denoted as data corpus "D" is passed to Level 4 for further processing,
b) Performing Level 4 Context-Based Processing, by which the data corpus "D" members selected during said step (4a) are processed so as to produce "context associations" using one or more of a variety of methods, which may be applied to either or both the elements of data corpus "D" or to additional databases and / or knowledge sources, such as are known to practitioners of the art, so as to extract refinement of both associations and concept classes as was described in said step (1a),

[c5] A method as claimed in c4 , further comprising the step of:
a) Performing Level 4 to Level 5 Transition, by which the "context associations" and / or context refinements found in said step (4b) are filtered by any one or more of various algorithmic means well known to the practition-

ers of this art so as to extract a subset of associations by application of one or more selection criteria, and the generality and meaning of this method is not dependent upon the specific nature of these criteria, and this subset denoted as data corpus "E" is passed to Level 5 for further processing,

b) Performing Level 5 Semantic-Based Processing, by which the data corpus "E" members selected during said step (5a) are processed so as to produce "semantic associations" and "semantic meaning and / or interpretation" using one or more of a variety of methods, such as are known to practitioners of the art, so as to extract further refinement of associations as was described in said steps (2b, 3b, and 4b), concept classes as was described in said step (1a), and additionally any knowledge-based and / or semantic-based information that can be associated with the elements of data corpus "E",

c) (Optionally) perform steps 5a and 5b as many times as necessary with defined processing unique to each step 5c and different from any previous step to define the apparatus to the number of levels desired.

[c6] A method as claimed in c1, c2, c3, c4 and / or c5,further comprising the step of:

Performing Level N to Level (N-X) Feedback Control, where "N" errors to any of Levels 2 through 5, and "X"

may take on any value from (1, ..., N-1) inclusive, by which one or more of the parameters governing any of the processes as described in said claims 1, 2, 3, and / or 4 are controlled by the feedback loop operating on outputs computed at Level N, where N > the controlled level (1, 2, 3, or 4), and where multiple feedback loops can be implemented in any given instantiation of this method,

[c7] method as claimed in c6, further comprising the step of: Performing a Utility Function computation and output, by which the "Feedback Loop" as described in said step (6) is modulated and controlled by means of a function so as to give either or both the user and / or an automated process the ability to control and "tune" the feedback loop so as to bring the overall system results to a desired level of performance, and where the formulation of said "Utility Function" follows he rules of practice as are well understood by practitioners of the art,

[c8] An apparatus for use with the processes described in said c1, the apparatus comprising: one or more data access and / or storage unit(s) "DS-1" coupled to receive and store as needed the data corpus "A", one or more computational processing unit(s) "CPU-1" coupled o receive the data corpus "A" and perform the processing as indicated in claim 1 "Level 1" processing, one or more

data storage unit(s) "DS-2" coupled to the computational processing unit "CPU-1" so as to receive and store the data corpus "B" that is generated as an output of the process described in said claim 1 "Level 1" processing,
one or more computational processing unit(s) "CPU-2" coupled to receive the data corpus "B" from "DS-2" and perform the processing as indicated in claim 1 "Level 2" processing,
one or more data storage unit(s) "DS-3" coupled to the computational processing unit "CPU-2" so as to receive and store the data corpus "C" that is generated as an output of the process described in said claim 1 "Level 2" processing,
a visualization and / or display unit or other means of providing viewing and / or results interpretation of either or both Level 1 and / or Level 2 processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c9] An apparatus as claimed in c8, wherein if said step (2) is employed as part of the method, then additionally there is:
one or more data access and / or storage unit(s) "DS-0" coupled to receive and store as needed the data corpus "A0", from stored and / or live data feeds, one or more computational processing unit(s) "CPU-0" coupled to re-

ceive the data corpus "A0" and perform the processing as indicated in claim 2 "Level 0" processing, and is for that purpose coupled to "DS-1" so that the outputs of the Level 0 processing can be stored and made available for Step (1),
(optionally) a visualization and / or display unit or other means of providing viewing and / or results interpretation of Level 0 processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c10] An apparatus as claimed in c8, or c9, wherein if said step (3) is employed as part of the method, then additionally there is:
one or more computational processing unit(s) "CPU-3" coupled to receive the data corpus "C" from "DS-3" and perform the processing as indicated in claim 3 "Level 3" processing, one or more data storage unit(s) "DS-4" coupled to the computational processing unit "CPU-3" so as to receive and store the data corpus "D" that is generated as an output of the process described in said claim 3 "Level 3"processing,
(optionally) one or more visualization and / or display unit(s) or other means of providing viewing and / or results interpretation of Level 3 processing, and / or making these results available to another process, whether

automated and /or semi-automated,

[c11] An apparatus as claimed in c10, wherein if said step (4) is employed as part of the method, then additionally there is:
one or more computational processing unit(s) "CPU-4" coupled to receive the data corpus "D" and perform the processing as indicated in claim 4 "Level 4" processing, and if more than one unit is so used, then appropriate coupling exists so as to transfer results between the processes as is necessary, one or more data storage unit(s) "DS-5" coupled to the computational processing unit "CPU-4" so as to receive and store the data corpus "E" thatis generated as an output of the process described in said claim 4 "Level 4" processing,
(optionally) one or more visualization and / or display unit(s) or other means of providing viewing and / or results interpretation of Level 4 processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c12] An apparatus as claimed in c11, wherein if said step (5) is employed as part of the method, then additionally there is:
one or more computational processing unit(s) "CPU-5" coupled to receive the data corpus "E" and perform the processing as indicated in claim 5 "Level 5" processing,

one or more data storage unit(s) "DS-6" coupled to the computational processing unit "CPU-5" so as to receive and store the data corpus "F" that is generated as an output of the process described in said claim 5 "Level 5" processing,
(optionally) a visualization and / or display unit or other means of providing viewing and / or results interpretation of Level 5 processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c13] An apparatus as claimed in c12, which additionally contains one or more computational and data storage units wherein the one or more "Feedback Loop(s)" as described in said step (6) are computed and stored, and which is (are) coupled to the appropriate Level N and Level (N-X) computational (CPU) units,
(optionally) a visualization and / or display unit or other means of providing viewing and / or results interpretation of Feedback Loop processing, and / or making these results available to another process, whether automated and /or semi-automated,

[c14] An apparatus as claimed in c13, which additionally contains one or more units wherein the one or more "Utility Function(s)" as described in said step (7) are computed, and which is (are) coupled to the appropriate "Feedback

Loop" computational (CPU) units,
(optionally) a visualization and / or display unit or other means of providing viewing and / or results interpretation of the one or more Utility Function(s), and / or making these results available to another process, whether automated and /or semi-automated,

[c15] An apparatus as claimed in c14, wherein the various units described in Claims (8) through (13) inclusive may be combined as appropriate for the purpose of enabling the processing and storage requirements as are needed to meet the stated purposes of Claims (1) through (7).

[c16] An apparatus as claimed in c15, wherein one or more of the various units and the processes which are supported by each unit or appropriate combination of data storage and computational processing units, is embodied as an existing tool, whether available as a research prototype or "commercial-off-the-shelf" implementation.